Paralelismo sintáctico-semántico para el tratamiento de elementos extrapuestos en textos no restringidos[1]

M. Saiz, P. Martínez-Barco, M. Palomar
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera San Vicente S/N. 03080 ALICANTE, España.
URL: http://gpl.dlsi.ua.es
{ max, patricio, mpalomar } @dlsi.ua.es

Palabras clave: Procesamiento de Lenguaje Natural, Extraposición a izquierdas, Gramáticas *Datalog*, Paralelismo sintáctico-semántico.

RESUMEN. *En este artículo presentamos un método basado en la teoría del paralelismo para la identificación y resolución de elementos extrapuestos en textos no restringidos. Esta teoría de paralelismo está basada en (Palomar 96) y se amplía con el desarrollo de técnicas de análisis parcial –en las que se estudia las partes relevantes del texto- que facilitan la resolución de los fenómenos lingüísticos.*
Nos basaremos en los programas Datalog extendidos (Dahl 94) (Dahl 95) como herramienta para la definición e implementación de gramáticas. Éstas no están basadas en reglas gramaticales sino en la detección de información relevante, relajando el proceso y ampliando el conjunto potencial de textos analizables.

1.- EL PROBLEMA.

El desarrollo de sistemas de procesamiento de lenguaje natural presenta dificultades específicas cuando se desea tratar con un conjunto amplio del lenguaje. Para ello, el principal problema que se aborda es la consecución del análisis de infinitas oraciones del lenguaje mediante mecanismos finitos. Sin embargo, este objetivo resulta computacionalmente difícil de tratar, por lo que, generalmente, se opta por acotar el lenguaje restringiéndolo a un subconjunto de oraciones (Moreno 93).
Cuando se requiere que el sistema de procesamiento automático de un subconjunto del lenguaje natural sea más ambicioso, es decir, que la cobertura del lenguaje sea lo suficientemente amplia para incluir oraciones complejas, entonces surge una serie de problemas adicionales que incrementan notablemente las dificultades de tratamiento (Palomar 96): extraposición a izquierdas, elipsis, anáfora, …
Algunas aproximaciones gramaticales, las más extendidas, utilizan los formalismos gramaticales lógicos para la construcción automática de estructuras sintácticas y semánticas proporcionando los mecanismos adecuados para el tratamiento de estos fenómenos. Entre estos formalismos podemos destacar las gramáticas *Datalog* (Dahl 94) (Dahl 95) y las gramáticas *Datalog Extendidas* (Moreno 97).
Ante las dificultades existentes en el tratamiento de estos fenómenos en un conjunto amplio del lenguaje, presentamos un método de análisis parcial para la resolución de la extraposición a

[1] Este artículo ha sido subvencionado por el proyecto CICYT nº TIC97-0671-C02-01/02

izquierdas en textos no restringidos. Nos basaremos, por una parte, en el paralelismo sintáctico-semántico (Palomar 96) para la identificación del elemento extrapuesto y, por otra, en la resolución semántica para la reconstrucción de la frase.
Siguiendo la definición clásica de F. Pereira (Pereira 81) la extraposición a izquierdas "ocurre en una oración cuando un subconstituyente de un constituyente que forma parte de la oración, se representa por otro a la izquierda del que está incompleto".
El problema de la extraposición a izquierdas en textos restringidos ha sido tratado en innumerables ocasiones. Sin embargo, estos métodos plantean importantes problemas para el tratamiento de este fenómeno ubicado en el marco de los textos no restringidos.
La siguiente frase representa un posible ejemplo de texto no restringido y que será analizada posteriormente:

Un teléfono es un aparato de telefonía que comprende, al menos, un transmisor telefónico, un receptor telefónico, el cableado y los órganos accesorios directamente asociados a estos transductores

Uno de los principales problemas en el planteamiento de los textos no restringidos, es la imposibilidad de contar con reglas gramaticales para analizar todos los casos posibles. La solución que planteamos se basa en la detección de la información relevante en el texto relajando las gramáticas con el fin de extender su campo de acción.
Para conseguir una aproximación al tratamiento y resolución de la extraposición a izquierdas, usaremos gramáticas Datalog Extendidas con un método de análisis basado en técnicas incrementales de evaluación aplicadas sobre estas gramáticas.
La implementación del algoritmo semi-naive (Dahl 94) aplicado a las gramáticas Datalog Extendidas (Moreno 97) nos permitirá "congelar" el proceso de análisis de la frase para aplicar el tratamiento metagramatical.

2.- ANTECEDENTES.

Existen numerosos trabajos en la literatura que hacen hincapié en la resolución de elementos extrapuestos, movimiento de elementos a larga distancia, extraposición a izquierdas, etc. Sin embargo, los trabajos que abordan la resolución de estos fenómenos normalmente acotan el lenguaje restringiéndolo a un subconjunto de oraciones.
Algunos de estos trabajos, los más destacados (Gramáticas de Extraposición (Pereira 81), Gramáticas Discontinuas (Dahl 88), Gramáticas Discontinuas Estáticas (Dahl 90), Gramáticas Discontinuas Restringidas (Dahl 86), Gramáticas de Movimiento de Literales (Groenink 95) pasan por ser una extensión de las Gramáticas de Cláusulas Definidas donde la aumentan y la extienden con nuevas reglas propias para el tratamiento de estos fenómenos lingüísticos.
Estos trabajos tienen el principal inconveniente en la necesidad de un análisis completo de las frases basándose en las reglas gramaticales, lo cual es prácticamente imposible si nos centramos en un gran volumen de textos no restringidos.
Por otro lado, las gramáticas *Datalog* (DLG) presentadas por V.Dahl, P.Tarau y Y.Huang en (Dahl 94) y ampliadas posteriormente por V.Dahl, L.Moreno, M.Palomar y P.Tarau en (Moreno 97) (Gramáticas *Datalog Extendidas)* están orientadas a la resolución de estos fenómenos lingüísticos en textos restringidos con sus problemas derivados.
Para solucionar estos problemas, planteamos una mejora en las Gramáticas *Datalog* Extendidas; utilizando técnicas de compilación apropiadas (p.ej. las técnicas incrementales), análisis parciales de los textos, y mediante un método basado en la teoría del paralelismo, podremos aportar soluciones a los problemas mencionados, tal y como veremos a continuación.

3.- EL PARALELISMO SINTÁCTICO-SEMÁNTICO EN LA RESOLUCIÓN DE EXTRAPOSICIÓN A IZQUIERDAS EN TEXTOS NO RESTRINGIDOS.

El método que presentamos en este artículo intenta resolver la extraposición basándose en el paralelismo sintáctico-semántico para identificar el elemento extrapuesto. Consideramos que dos estructuras son paralelas si son isomórficas y tienen las mismas condiciones sintácticas y semánticas. En concreto, para el caso de la extraposición, diremos que el elemento extrapuesto y la traza son paralelos sintáctica y semánticamente si el elemento extrapuesto cumple las condiciones esperadas por la traza o el hueco dejado.

La idea principal que subyace en este nuevo método es el uso de la resolución semántica en la reconstrucción de la frase en lugar de usar las aproximaciones sintácticas. Así, la forma lógica del antecedente se identifica con el elemento extrapuesto. Cuando se producen problemas de ambigüedad, se utiliza un mecanismo de tipo semántico que pueda resolver un elevado número de casos.

El tratamiento del fenómeno lingüístico de la extraposición de elementos, necesita identificar en primer lugar cuáles son los constituyentes que han sido extrapuestos y cuál es su nueva posición, y en segundo lugar, cuál es la posición que ha quedado vacía en la oración de relativo (traza).

3.1. Análisis mediante técnicas incrementales.

El método está basado en el algoritmo *semi-naive* (Dahl 94). En este algoritmo se parte de un conjunto de axiomas y, mediante la aplicación de reglas de derivación, se obtienen los teoremas del primer nivel. Estos teoremas se tomarán como un nuevo punto de partida de tal forma que mediante la aplicación de las reglas de derivación, se obtendrán los teoremas del segundo nivel, y así sucesivamente. Generalmente, para derivar los teoremas de un nivel, se deberá usar al menos uno de los teoremas producidos en el nivel anterior. El proceso termina cuando ya no se pueden generar más teoremas.

El paralelismo sintáctico-semántico nos ayudará a determinar las estructuras paralelas de forma automática (el elemento extrapuesto y la traza), mediante la aplicación incremental de una restricción de gramática Datalog con predicción top-down que completará la estructura perdida a través de un análisis de paralelismo inspirado en (Palomar 96).

En el algoritmo *semi-naive* comprobaremos en cada paso de la derivación incremental de teoremas si se ha derivado un teorema de la forma *C(prel, M, N)*. Cuando esto ocurre, se añade en este mismo momento una restricción para detectar la ausencia (traza) de un sintagma nominal en la oración de relativo y para relacionar sintáctica, morfológica y semánticamente dicha traza con algún constituyente anterior (el elemento extrapuesto).

3.2. Definición del Algoritmo.

Las fases del algoritmo que proponemos son:

a) Detección de la existencia de un pronombre de relativo.
b) Identificación de información relevante: sintagmas nominales, preposicionales y verbales.
c) Determinación de la traza y del elemento extrapuesto en la oración mediante la aplicación de las restricciones sintácticas, morfológicas y semánticas adecuadas.
d) Reconstrucción semántica de la oración

Veamos el siguiente ejemplo con las posiciones que ocupa cada palabra en la frase:

$_0$ El $_1$ perro $_2$ que $_3$ tenía $_4$ la $_5$ rabia $_6$ mordió $_7$ a $_8$ la $_9$ mujer $_{10}$

La figura muestra un seguimiento paso a paso aplicando la técnica al ejemplo anterior:

T1 = {C(El,0,1) , C(perro,1,2) , C(que,2,3) , C(tenía,3,4) , C(la,4,5) , C(rabia,5,6) , C(mordió,6,7) , C(a,7,8) , C(la,8,9) , C(mujer,9,10)}

T2 = T1 union {
C (d (sem (X , Y , Z , existe (X , Y , Z))), sin (m , s) , res (R) , 0 , 1) ,
C (n (sem (X , perro (X)) , sin (m , s) , res ([animal]) , 1 , 2) ,
C (prel (sem (X , Y , X & Y)) , sin (G , N) , res (R) , 2 , 3) ,
C (v (sem (X , Y , tener (X , Y))) , sin (G , s) , res (R) , 3 , 4) ,
C (d (sem (X , Y, Z , existe (X , Y , Z))), sin (f , s) , res (R) , 4 , 5) ,
C (n (sem (X , rabia (X) , sin (f , s) , res ([enfermedad]) , 5 , 6) ,
C (v (sem (X , Y , morder (X , Y))) , sin (G , s) , res (R) , 6 , 7) ,
C (prep (sem (X , Y , a (X , Y))) , sin (G , N) , res (R) , 7 , 8) ,
C (d (sem (X , Y , Z , existe (X , Y , Z))), sin (f , s) , res (R) , 8 , 9) ,
C (n (sem (X , mujer (X)) , sin (f , s) , res ([persona]) , 9 , 10)
}

T3 = T2 union {
C (sn (sem (X , Y , exist (X , perro (X))) , sin (m , s) , res ([animal]) , 0 , 2) ,
C (sn (sem (X , Y , exist (X , rabia (X))) , sin (f , s) , res ([enfermedad]) , 4 , 6) ,
C (sn (sem (X , Y , exist (X , mujer (X))) , sin (f , s) , res ([persona]) , 8 , 10)
}

T4 = T3 union {
C (sv (sem (X , existe (Y , rabia (Y) , tener (X , Y))) , sin (G , s) , res (R) , 3 , 6) ,
C (sp (sem (X , Y , existe (X , mujer (X) , a (X , Y))) ,
sin (f , s) , res ([persona]) , 7 , 10)
}

En este punto, se puede detectar el elemento extrapuesto (en este caso, las restricciones sintácticas, semánticas y morfológicas han detectado la ausencia del sujeto de la oración determinando que el elemento extrapuesto es “el perro”) y se reconstruye la oración de relativo añadiendo su información semántica. Así podemos postular:

C (o (sem (existe (X , perro (X) , existe (Y , rabia (Y) , tener (X , Y))))) ,
sin (_ , _) , res (_) , 3 , 6)

T5= T4 union {
C(orel (sem (X , existe (X , perro (X) , existe (Y ,rabia (Y), tener (X , Y))))),
sin (_ , _) , res (_) , 2 , 6)
C (sv (sem (X , existe (Y , mujer (Y) , morder (X , Y)))) ,
sin (G , s) , res (R) , 6 , 10) ,
}

T6=T5 union {
C (sn (sem (X , Z , existe (X , perro (X) & existe (Y , rabia (Y) , tener (X , Y)) , Z))),
sin (m , s) , res ([animal]) , 0 , 6)
}

T7=T6 union {

C (o (sem (existe (X , perro (X) & existe (Y , rabia (Y) , tener (X , Y)) ,
existe (Z , mujer (Z) , morder (X , Z))))) , sin (_ , _) , res (_) , 0 , 10)
}

Puede comprobarse que al finalizar el análisis, se han detectado las posiciones inicial y final (0 y 10) obteniendo así la forma lógica de la frase completa.
Nótese que mientras en las anteriores propuestas para la resolución del problema de la extraposición a izquierdas basadas en lógica, como las Gramáticas de Extraposición (Pereira 81) y las Gramáticas Discontinuas (Dahl 88), se hacía una reconstrucción sintáctica donde el antecedente se reescribía como el componente perdido, nuestro análisis con reconstrucción semántica se aproxima más a la comprensión humana.

3.3. Posibles problemas y sus soluciones.

Podemos encontrarnos con casos como el que sucede en la siguiente oración:

El soldado de plomo que dispara al aire pertenece al Ejército de Infantería.

En este caso, el algoritmo podría detectar que el elemento extrapuesto es el sintagma nominal extrapuesto es plomo ya que cumple las restricciones anteriormente expuestas;
El análisis semántico posterior determinaría que semánticamente no es posible que concuerde el elemento extrapuesto (sujeto del verbo) con el sintagma nominal plomo, ya que disparar no es un rasgo semántico de éste. Por lo tanto el análisis devolvería fallo. Aunque no es el caso que nos ocupa, la concordancia en el análisis morfológico también nos puede ayudar a resolver la ambigüedad. Esto ocurre en la siguiente frase:

Los árboles del bosque que rodean la casa dejan caer sus hojas en el otoño.

3.4. Desarrollo del paralelismo para textos no restringidos.

Una vez aclarado el problema de la extraposición a izquierdas, y su resolución con el uso de Gramáticas Datalog Extendidas, nos centraremos en el tratamiento de los textos no restringidos. Con el fin de abarcar este tipo de textos sin contar con reglas universales, obviaremos en el análisis aquella información que no consideremos relevante. Esta información no relevante constará de aquellos elementos sintácticos que no resulten ser sintagmas nominales, sintagmas verbales y sintagmas preposicionales.
Tomaremos como base un corpus con cinco millones de palabras etiquetadas automáticamente por la versión española del etiquetador de Xerox[2]. Este corpus contiene el manual de la *International Telecommunications Union CCITT*, también conocido como *The Blue Book*, en su versión en castellano y supone la colección más importante de textos de telecomunicaciones Además de la sintaxis, la etiqueta de cada palabra también contiene su información morfológica (número, género y persona).
El procedimiento para la resolución del elemento extrapuesto en textos no restringidos será el siguiente:

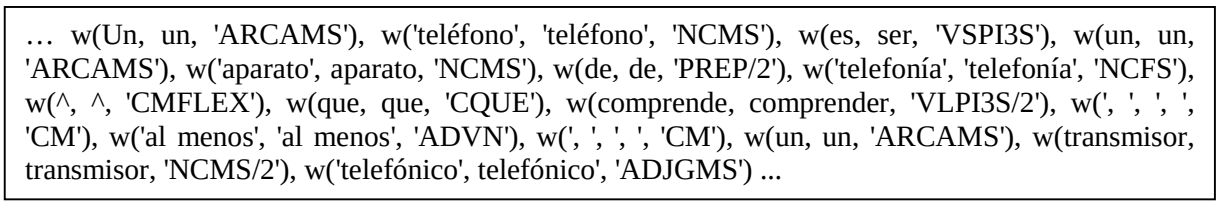
[2] La versión española del etiquetador Xerox pertenece al proyecto CRATER (**C**orpus **R**esources **A**nd **T**erminology **E**xt**R**action). http://lola.lllf.uam.es/~fernando/projects/es_corpus.html

- Trabajaremos directamente sobre la salida del corpus etiquetado, realizando posteriormente un análisis parcial ascendente del texto, en el que se identificarán los sintagmas nominales, sintagmas preposicionales, sintagmas verbales y pronombre relativo.
- Para cada una de las frases, en el proceso de análisis y una vez detectado el pronombre relativo, se aplicará el método del paralelismo anteriormente presentado para detectar el movimiento de elementos y poder llevar a cabo su reconstrucción semántica. Para este proceso, la información que contenga el verbo será fundamental, ya que tomando como base los huecos de objetos llenados de dicho verbo, podremos obtener los huecos no llenados que serán susceptibles de haber sido extrapuestos.

Veamos la aplicación de este procedimiento en la siguiente oración ejemplo:

Un teléfono es un aparato de telefonía que comprende, al menos, un transmisor telefónico, un receptor telefónico, el cableado y los órganos accesorios directamente asociados a estos transductores.

La siguiente figura muestra un extracto de la salida que genera el etiquetador sobre la frase anterior, y que tomamos como base para nuestro análisis:

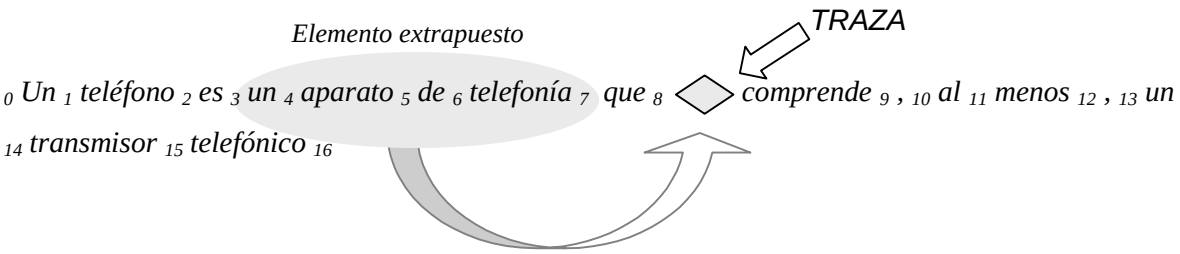


Partiendo de la frase etiquetada, y una vez aplicado el algoritmo semi-naive, trabajaremos únicamente con la información que consideramos relevante, es decir, los sintagmas nominales, verbales, preposiciones y el pronombre relativo.
La siguiente tabla muestra dicha información extraída del ejemplo en los distintos pasos del algoritmo. Aunque la información extraída es tanto sintáctica como semántica y morfológica, anotaremos únicamente, con el fin de simplificar, los componentes sintácticos:

$_0$ Un $_1$ teléfono $_2$ es $_3$ un $_4$ aparato $_5$ de $_6$ telefonía $_7$ que $_8$ comprende $_9$, $_{10}$ al $_{11}$ menos $_{12}$, $_{13}$ un $_{14}$ transmisor $_{15}$ telefónico $_{16}$

Pasos	*Sint. nominales*	*Sint. verbales*	*Sint. preposicionales*	*Pro. relativo*
1	C(sn,0,2) C(sn,3,5) C(sn,6,7) C(sn,13,16)		C(sp,5,7)	C(prel,7,8)
2	C(sn,3,7)	C(sv,8,16)		

En ese punto, ya se ha detectado el pronombre relativo, la traza y el elemento extrapuesto. Tras efectuar la reconstrucción de la oración de relativo y resolver la extraposición, se continuaría analizando como en el ejemplo visto en 3.2.



Puede verse como los elementos no relevantes de la oración (*p.ej. "al menos,"*) han sido deliberadamente omitidos.

4.- CONCLUSIONES.

Hemos presentado un método basado en la teoría del paralelismo para la identificación de elementos extrapuestos o movimiento de elementos a larga distancia en textos no restringidos. Para ello hemos realizado análisis ascendentes parciales para la detección de información relevante en el texto. Este análisis está desarrollado según las técnicas incrementales de las gramáticas *Datalog extendidas*, que han sido aumentadas y extendidas incluyendo la posibilidad de contar con elementos desconocidos irrelevantes para la resolución de estos fenómenos lingüísticos.

REFERENCIAS.

Dahl, V. (1988): *Discontinuous Grammars*. Informe interno CSS/LCCR 88-26. Simon Fraser University. Burbany, Canada 1988

Dahl, V. y Popowich, F. (1990): *Parsing and Generation with Static Discontinuity Grammars*. New Generation Computing, 8. 1990

Dahl, V. (1986): Saint-Dizier, P. *Constrained Discontinuous Grammars a linguistically motivated tool for processing language*. TR LCCR 86-11, Simon Fraser University, 1986

Dahl, V., Tarau P. y Huang Y (1994): *Datalog Grammars*. Proc. of the GULP-PRODE'94. V. 2. ed. UPV, 1994

Dahl, V., Tarau P., Moreno, L. y Palomar, M. (1995): "Treating coordination with Datalog Grammars". Computational Logic for natural Language Processing. (CLNLP-95). Edimburgo. Escocia. 1995

Groenink, A. (1995): *Literal Movement Grammars*. 1995.

Moreno, L. (1993): *Formalismos lógicos para el análisis e interpretación oracional del lenguaje natural*. Tesis doctoral. (FI- Universidad Politécnica de Valencia)

Moreno, L., Palomar, M. y Molina A. (1997): *Gramáticas Datalog Extendidas: una nueva aproximación*. APPIA-GULP-PRODE'97. Join Conference on Declarative Programming. Grado. Italia. Junio 1997.

Palomar, M. (1996): *Aportaciones a la resolución de la elipsis en lenguaje natural utilizando técnicas incrementales.* Tesis doctoral. (FI- Universidad Politécnica de Valencia)

Pereira, F.C.N. (1981): *Extraposition Grammars*. Computational Linguistics, 7(4), 1981.